Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 1 (2012) 464 – 468

Procedia Technology

INSODE-2011

Automatic extraction of mathematical terms for precalculus

Velislava Stoykova[a]*, Ekaterina Petkova[b]

[ab]*Institute for Bulgarian Language – BAS, 52, Shipchensky proh. str, bl. 17, 1113 Sofia, Bulgaria*

Abstract

In this work, we present the results of research for evaluating a methodology for extracting mathematical terms for precalculus using the techniques for semantically-oriented statistical search. We use the corpus-based approach and the combination of different statistically-based techniques for extracting keywords, collocations and co-occurrences incorporated in the Sketch Engine software. We evaluate the collocations candidate terms for the basic concept *function(s)* and approve the related methodology by precalculus domain conceptual terms definitions. Finally, we offer a conceptual terms hierarchical representation and discuss the results with respect to their possible applications.

Keywords: Artificial Intelligence; Information technology & Languages; Knowledge Engineering; Automatic Term Extraction.

1. Introduction

Recent development in information technologies offers artificial intelligence (AI) based methodologies for almost all fields including terminology and can result in a wide range of real applications. The modern lexicography and terminology uses extensively the electronic text corpora of different genres instead text archives used in the past. Existing electronic text collections of different types increase the necessity of proper software tools assisting different types of search to make the best use of them.

2. Computer-assisted terminology extraction

The problem of optimizing searching and finding the word context is a central for word sense definitions in its traditional interpretation in terminology frameworks but it can be, also, addressed as a process of creating terminological conceptual knowledge hierarchy which defines the semantic terminological relations (in AI frameworks). Further, we are going to present and discuss the results of a corpus-based research and analysis of the web-based open-source mathematical texts for precalculus given at the Wikipedia for computer-supported extraction of conceptual terminological database design.

The computer-assisted terminology extraction has been the most successful technique recently developed and applied for the creation of highly structured and semantically-oriented lexical reference sources like dictionaries, thesauri, etc. It is based on the extensive use of electronic text corpora assisting various types of search procedures. There are two general types of corpus-based applications – using rule-based search techniques (mostly by encoding

* Velislava Stoykova. Tel.: 359 2 979 2953; fax: +359 2 970 23 02.
E-mail address: *vili1@bas.bg*.

doi:10.1016/j.protcy.2012.02.102

grammatical relations like inflection [1], syntax, etc.) and using statistically-based search techniques (mostly by extraction of domain conceptual semantic relations as presented in [2] and [3]). But both are focused on the investigation of word behavior in the large-scale electronic text collections by using the related context definitions.

Further, we are going to present the methodology for term extraction based on the combination of different statistical search techniques to define the semantics of some general mathematical terms for precalculus based on the use of a corpus-based approach allowed by the Sketch Engine software.

3. The Sketch Engine's approaches to term extraction

The Sketch Engine software [4] allows the use of various approaches to term extraction and most of them might be for multilingual application. Generally, they are of two types – grammatically-based (using the related inflected word forms or part-of-speech categorization frames for measuring semantic similarity) and statistically-based (using different statistical approaches to define terms and their conceptual semantic relations). It is possible also to use the combinations of both approaches, however, for our research, we have used only statistically-based functions incorporated in the Sketch Engine software.

Extracting keywords is a most common and widely used technique to define the basic terms of a particular domain. The Sketch Engine's software standard options for keywords are based on the use of word frequency lists evaluation. However, in this way, it is possible to evaluate mostly the general basic terms of a particular domain and the application of additional statistical techniques is needed to define the basic terms' semantic conceptual relations.

Generally, semantic conceptual relations are regarded as to be of two types - horizontal and vertical. The horizontal semantic relations are those of synonymy, anthonymy, meronymy, i. e. showing semantic similarity [3] or semantic distance. The vertical semantic relations express the semantics of ordering or hierarchy and are realized by hyperonymy and hyponymy. All types of semantic relations can be extracted by generating the related word contexts through the generation of related word concordances based on the use of different statistical corpus-based approaches [2].

The concordances give all occurrences of the word in its related contexts and can be generated by using statistical search [3]. The Sketch Engine software standard options for concordance generation are flexible and can be expanded with respect to the related amount of words before and after the target keyword. Concordances define context in quantitative terms and a further work is needed to be done to define the semantic relations by searching for co-occurrences and collocations of the related keyword.

Co-occurrences and collocations are words which are most probably to be found with the related keyword. They assign the semantic relations between the keyword and its particular collocated word which might be of similarity or of a distance. The statistical approaches we are using to search the co-occurrence and collocated words are based on defining the probability of their co-occurrences and collocations. We have used the techniques of *T−score*, *MI−score* [5] and MI^3 *− score* [2] incorporated in the Sketch Engine for corpus processing and searching.

Basically for all, the following terms are used: N - corpus size, f_A - number of occurrences of the keyword in the whole corpus (the size of the concordance), f_B - number of occurrences of the collocated keyword in the whole corpus, f_{AB} - number of occurrences of the collocate in the concordance (number of co-occurrences). The related formulas for defining *T−score*, *MI−score* and MI^3−score are as follows:

$$\text{MI-Score} \quad \log_2 \frac{f_{AB} N}{f_A f_B}$$

$$\text{T-Score} \quad \frac{f_{AB} - \frac{f_A f_B}{N}}{\sqrt{f_{AB}}}$$

$$\text{MI}^3\text{-Score} \quad \log_2 \frac{f_{AB}^3 N}{f_A f_B}$$

T−score, *MI−score* and MI^3−score use different statistical criteria to evaluate co-occurrences and collocation words of a particular keyword, however for our research, we have used *T−score* criterion for ranging, even the results for *MI−score* and MI^3−score criteria are listed as well.

4. Extracting terms by using statistical search

Traditionally, terms extraction and definition of their conceptual relations are based mostly on expertise of different contexts of a particular word (a possible term). However, the context can be defined in logical or linguistic terms depending on the theory which is used. For our research, we use the context in quantitative statistical terms since our aim is to extract the precise term meaning and its relations by using as much as possible related contexts.

4.1. Extracting terms by using keywords frequency lists

We have created the electronic text corpora MathWiki consisting of texts for precalculus given at the Wikipedia [6] and mostly based on mathematical concepts' relations interpretation introduced in [7] of almost 150 000 words. Also, we use the British National Corpus (BNC) to compare the research results for term extraction. There are various statistical approaches to define keywords. In general, most of them define the task for keywords extraction as the retrieval and clustering of statistically similar words [8] and they differ with respect to the statistics used. However, we use the statistics incorporated in the Sketch Engine ([4], [9]) software for processing corpora which allows the use of combined statistical approaches and the comparison of the results between several corpora [10].

After using the Sketch Engine's standard statistical options for processing our text corpus for keywords definition, we have obtained the results given at Fig. 1. The results reflect the encyclopaedic knowledge presentation of MathWiki corpus and represent the basic precalculus concepts defined in [11] like *function(s)*, *numbers*, *polynomials*, *graphs*, *equations*, etc.

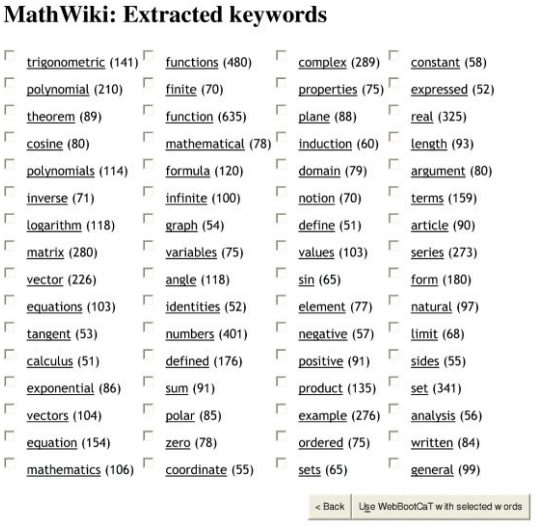

Fig. 1. Extracted keywords from MathWiki corpus.

Thus, our keywords frequency list give the general mathematical concept terms of precalculus, however, the proportion and their order have to be clarified by using further statistical search to define their semantic relations. In further description, we are going to define the semantic relations only for the basic concept term *function(s)*.

4.2. Generating keywords concordances

Currently, the statistical corpus approaches based on the measurement of words similarity and defining words concordances have been widely used in terminology for conceptual term definition. The concordances give all occurrences of a word in its related contexts from all available text sources and can be generated by using statistical search [3]. The example concordances for the basic concept term *function* received from MathWiki corpus is given at Fig. 2.

Corpus: MathWiki

Page 1 of 33 Go

Next | Last

initiated by Descartes. A function	, in mathematics, associates one
the argument of the function	, also known as the input, with
quantity, the value of the function	, also known as the output.
known as the output. A function	assigns exactly one output to
given set. An example of a function	is f(x) = 2x, a function which
a function is f(x) = 2x, a function	which associates with every
f(5) = 10. The input to a function	need not be a number, it can
object. For example, a function	might associate the letter A with
describe or represent a function	, such as a formula or algorithm

Fig. 2. Generated concordances for concept term *function* from MathWiki corpus.

The received concordances consist of 33 pages context results which can be used for conceptual term definitions or for further semantic filtering by generating collocations or co-occurrences of a related keyword *function*.

4.3. Extracting conceptual semantic relations by using co-occurrences and collocations

The related corpus query systems allow great flexibility of statistically-based search including also further semantic filtering and refinement by generating co-occurrences and collocations using different statistical techniques where the context is defined in quantitative terms. The most likely collocations candidate words for MathWiki corpus (which are the most frequent collocates) for the keyword *function* are given at Fig. 3 and are *exponential*, *rational*, *polynomial*, *complex*, *logarithmic*, *trigonometric*, etc. They express the hierarchical conceptual semantic relations of the keyword reflecting in their ranging.

Collocation candidates MathWiki

Page 1 Go

Next >		Freq	T-score	MI	MI3
p/n	exponential	65	8.001	7.040	19.085
p/n	inverse	29	5.309	6.152	15.868
p/n	rational	24	4.804	5.689	14.859
p/n	propositional	20	4.437	7.011	15.655
p/n	polynomial	23	4.544	4.253	13.301
p/n	complex	19	3.978	3.517	12.013
p/n	logarithm	12	3.268	4.146	11.316
p/n	trigonometric	11	3.072	3.764	10.683
p/n	increasing	8	2.794	6.357	12.357
p/n	tangent	8	2.721	4.716	10.716
p/n	relation	8	2.719	4.689	10.689
p/n	continuous	7	2.605	6.004	11.618

Fig. 3. Collocations candidates for keyword *function* for MathWiki corpus ranged according to *T−score* criterion.

Alternatively, the relatively not too frequent collocations like *periodic*, *continuous*, *inverse*, *increasing*, *decreasing*, *real-valued*, *multi-valued*, *positive*, *negative*, etc. represent the attributive semantic relationships of the keyword *function*.

4.4. Building conceptual semantic hierarchy

The most frequent collocations represent the relationship of similarity but they do not necessarily express the semantic relationship of synonymy. Generally, the conceptual semantic term relations extracted by collocations and co-occurrences mostly represent vertical semantic relations like hyponymy or hyperonymy. Thus, for our research results, we are using such interpretation and we define *polynomial function*, *exponential function*, and *rational function* as the most important hyponymic concepts of the very general hyperonym conceptual term *complex function*. The *logarithmic function* can be presented as the inverse to the *exponential function* and the *trigonometric function* can be presented as divided into its subsequent parts *sine*, *cosine*, *tangent*, and *cotangent functions*. The semantic conceptual hierarchy of the basic concept term *function* is presented at Fig. 4.

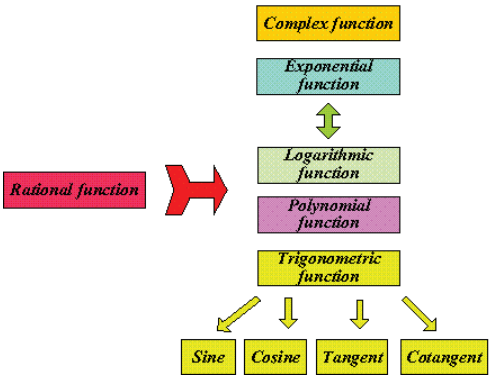

Fig. 4. Conceptual terminological hierarchy of the concept term *function*.

5. Conclusion

The extracted terms and semantic relations show that statistically-based search technique for extracting keywords, collocations and co-occurrence words is effective approach for mathematical conceptual precalculus terms extraction from specialised web-corpus. The terms are evaluated on the base of their high frequency in the MathWiki corpus and their relatively low frequency in the BNC. In fact, the term *precalculus* was not occurred in the BNC. The evaluated methodology may be applied for fast production of up-to-date terminological reference sources (like specialized dictionaries or thesauri) or building ontologies (for defining the logical relations, conceptual relations or hierarchies).

References

1. V. Stoykova, Bulgarian noun – definite article in DATR. In D. Scott (ed.) *Artificial Intelligence: Methodology, Systems, and Applications. Lecture Notes in Artificial Intelligence 2443*, Springer-Verlag, (2002), pp.152–161.

2. M. Oakes, *Statistics for Corpus Linguistics.* Edinburgh University Press, (1998).

3. K. Sparck Jones, *Synonymy and Semantic Classification.* Edinburgh University Press, (1986).

4. A. Kilgarriff, P. Rychly, P. Smrz and D. Tugwell, The Sketch Engine. In *Proceedings from EURALEX 2004.* Lorient, France, (2004), pp. 105–116.

5. K.Church and P.Hanks, Word Association Norms, Mutual Information and Lexicography.*Computational Linguistics* 16:1, (1991), pp.22-9.

6. Precalculus, (2011), `http://en.wikipedia.org/wiki/Precalculus`

7. M. Hazewinkel, *Encyclopaedia of Mathematics,* Springer-Verlag, (2001).

8. Dekang Lin, Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the COLING-ACL.* Montreal, (2002), pp. 768-774.

9. A. Kilgarriff and M. Rundell, Lexical Profiling Software and its Lexicographic Applications: a Case Study. In *Proceedings from EURALEX 2002.* Copenhagen, (2002), pp. 807-811.

10. A. Kilgarriff, S. Reddy, J. Pomikalek and P. Avinesh, A Corpus Factory for Many Languages. In N. Calzolari (ed.) *Proceedings of the LREC 2010*, Malta, (2010), pp. 904-910.

11. V. Stoykova and M. Mitkova, Defining Lexical Semantic Relationships for Terms of Precalculus Study. In N. Mastorakis, V. Mladenov, and Z. Bojkovic (eds.) *Recent Researches in Educational Technologies*, Corfu, Greece, (2011), pp. 240-244.